Regular paper | Journal of the KIECS. pp. 289-294, vol. 19, no. 1, Feb. 29. 2024, t. 123, pISSN 1975-8170 | eISSN 2288-2189
http://dx.doi.org/10.13067/JKIECS.2024.19.1.289

감정 분석 기반의 선호도 분석 시스템의 설계 및 구현

문희준 · 김동현*

Design and Implementation of A Preference Analysis System Based on Sentiment Analysis

Hee-Jun Moon · Dong-Hyun Kim*

요 약

전통적인 여론 조사 기반의 선호도 분석 기법은 많은 시간과 비용이 요구되고 조사할 수 있는 도메인이 제한적인 문제가 있다. 이를 해결하기 위하여 이 논문에서는 감정 분석 기반의 선호도 분석 시스템을 제안한다. 사용자가 입력한 키워드를 이용하여 웹 문서를 수집한 후에 N-gram 기법을 이용하여 극성을 계산한다. 다량의 웹 문서를 분석할 때 발생하는 분석 시간을 줄이기 위하여 워커 서비스를 사용하는 컨테이너 기반의 시스템을 설계하고 구현하였다. 제안 시스템의 분석 결과와 기존 여론 조사를 비교하였을 때 1% ~ 8%의 오차를 보여준다.

ABSTRACT

Traditional poll-based preference analysis techniques are time-consuming, expensive, and limited in the domains they can survey. To solve this problem, this paper proposes a preference analysis system based on sentiment analysis. After collecting web documents using the keywords entered by the user, the polarity is calculated using the N-gram technique. To reduce the analysis time when analyzing a large amount of web documents, we designed and implemented a container-based system using worker services. Comparing the analyzed results of the proposed system with existing polls shows a difference of 1% to 8%.

키워드

Sentiment Analysis, Preference, Poll, Worker Service, Container
감정 분석, 선호도, 여론 조사, 워커 서비스, 컨테이너

I. 서 론

최근 다양한 매체와 인터넷의 발달로 많은 양의 정보가 생성과 동시에 소비되고 있으며 다양한 정보를 기반으로 분석을 통하여 새로운 정보를 생산하는 요구와 시도는 계속하여 증가하고 있다. 이러한 시도의 대표적인 예로 여론 분석을 통한 선호도 예측이 있다[1]. 기존의 선호도 예측에서는 여러 가지 기법이 존재하나 전통적인 방법으로는 전화 또는 설문을 통한 여론조사가 있다. 그러나 전통적인 여론조사는 많은 비용 소모가 요구되고 조사하는 스펙트럼이 한정적인 단점이 존재한다. 이를 보완하기 위하여 소셜네트워크 또는 매체 데이터를 이용한 분석 시도가 이루어지고 있으며 대표적으로 구글에서 제공하는 구글 트렌드가 있다. 구

동서대학교 소프트웨어학과(aglide100@gmail.com)
* 교신저자 : 동서대학교 소프트웨어학과
· 접 수 일 : 2023. 11. 01
· 수정완료일 : 2023. 12. 22
· 게재확정일 : 2024. 02. 17
· Received : Nov. 01, 2023, Revised : Dec. 22, 2023, Accepted : Feb. 17, 2024
· Corresponding Author : Dong Hyun Kim
Dept. of Software, Dongseo University,
Email : pusrover@dongseo.ac.kr

글 트렌드는 특정 제품 또는 인물에 대하여 일정 기간 동안의 검색량을 측정한 후에 검색량 기반의 호감도를 분석한 결과를 제공한다. 그러나 구글 트렌드는 단순한 검색량만을 통해 분석하기 때문에 호감도 측정을 하면 오류를 내재할 가능성이 있다. 따라서 검색량뿐만 아니라 문장이 가지는 감정을 측정할 필요가 있다.

감정 분석을 활용한 선행 연구로 한국어 감성 분류에 적합한 모델을 찾기 위하여 RNN 기법과 트랜스포머 계열 파생 기법들을 비교 평가하였다[2]. 다양한 입력 자료를 이용하여 감정을 분석하고 표정과 음성과의 유사도를 측정하기 위한 인공지능 모델이 제안되었으며[3], 소셜네트워크에서의 감정을 데이터 시각화 기법을 통하여 시각적으로 표현한 기존 사례들을 분석하였다[4]. 또한, 감정 사전을 기반으로 통합 벡터를 계산하는 감정 분석 기법을 제시하였으며[5], 감정 분석을 통하여 적합한 캐릭터를 생성 및 디스플레이하는 AI스피커가 제시되었다[6]. 또한 사회적 재난에 대한 시민 감성도를 트위터 데이터를 활용하여 분석한 결과를 제시하였다[7].

본 논문에서는 감정 측정을 이용한 선호도 분석 시스템을 제안한다. 제안 시스템은 사용자가 입력한 키워드를 기반으로 검색 엔진을 통하여 웹 문서데이터를 추출하고 추출된 웹 문서데이터에 대한 감정 측정을 통해 입력 키워드에 대한 호감도를 평가한다. 감정 측정을 위하여 사전에 구축된 감정 사전과 형태소 기반의 N-gram 기법을 사용하여 극성을 분석한다. 대용량 웹 데이터에 대하여 극성 분석을 할 경우 발생하는 분석 시간 소요 문제를 해결하기 위하여 워커 서비스를 사용하는 컨테이너 기반의 시스템을 설계하고 구현한다.

본 논문의 구성은 다음과 같다. 2장에서 관련 연구를 기술하고 3장에서 키워드 기반의 선호도 분석 시스템의 감정 측정 및 설계를 제안한다. 4장에서 시스템 구현 결과에 관하여 기술하고 마지막으로 5장에서는 결론을 기술한다.

II. 관련 연구

한국어 감성 분류를 위한 RNN 기법과 트랜스포머 기법을 비교평가하였다[2]. 이를 위하여 네이버의 영화 및 쇼핑 리뷰 자료, 게임 사용 리뷰 자료를 수집하고 감성 분류에 사용될 수 있는 RNN과 트랜스포머 파생 모델인 BERT와 GPT를 사용하여 정확도를 비교하였다. 실험 결과로써 RNN 기반의 딥러닝 모델보다는 트랜스포머 계열 모델들이 우수한 성능을 보였으며, 특히 BERT 계열의 LMKor-BERT가 상대적으로 제일 우수한 정확도를 보였다.

다양한 입력 자료를 이용하여 감정을 분석하고 표정 연습을 도와주기 위한 인공지능 모델이 제안되었다[3]. Open CV를 이용하여 얼굴 관심 영역을 추출하고 CNN을 기반으로 표정의 감정 수치를 계산하였다. 수치 계산을 위한 자료로는 FER2013 공개 데이터를 활용하였다. 또한 두 번째 입력 자료로 자연어 문장을 KoBERT모델을 사용하여 감정 수치를 계산하였다. 두 개의 입력 자료는 코사인 유사도를 이용하여 표정에서 측정된 감정과 자연어 문장에서 특정된 감정 자료의 유사도를 계산하였다.

소셜 네트워크에서의 감정을 시각적 표현 방식을 이용하여 표현한 사례 연구들을 분석하였다[4]. 이를 위하여 데이터에서 패턴 분석을 통하여 도출해 낸 의미들을 데이터 시각화를 통하여 표현한 기존 연구들을 분석하였으며 트위터 문서 분석을 통한 주가 예측, 트위터 영화 자료에 대한 감정 분석을 통한 방사형 그래프 표현, 인스타그램의 감정 형용사 분석 등을 살펴보았다. 그리고 이를 통해 감정 분석의 시각화를 위해 활용할 수 있는 요소들을 도출하였다.

감정 사전을 기반으로 통합 벡터를 계산하는 감성 분석 기법을 제안하였다[5]. 감성 사전을 통한 극성 추출 모델을 사용하여 단어의 감성 극성과 문장의 감성 극성을 추출하였다. 또한 인공신경망을 이용하여 특정 도메인의 문맥 내 감성 극성을 추출하여 3가지 값을 혼합한 벡터를 계산하였다. 또한 비교 실험을 통하여 단어 극성만 사용하였을 때와 문장 극성을 혼합하였을 때를 비교하여 문장 극성이 성능 향상에 많은 기여를 하였음을 보였다.

감정 분석을 통하여 캐릭터를 디스플레이하는 AI 스피커를 제시하였다[6]. 독거노인의 움직임과 음성을 라즈베리파이기반의 센서와 마이크를 통하여 입력받아 파일로 생성하였고, 이를 관제 서버에 전달하였다. 관제 서버는 구글스피치API를 이용하여 텍스트 자료로 변환 후에 DialogFlow 서버로 전송하여 감정 분석

을 기반으로 상황에 맞게 캐릭터를 생성하였다. 또한, 사회적 재난에 대한 시민 감성도 분석 기법이 제안되었다[7]. 이 연구에서는 시민 건강에 경각심을 주었던 '옥시' 사건과 시민들에게 정신적 불안감을 주었던 '묻지 마 범죄' 사건에 대하여 트위터 데이터를 수집하였다. 그리고 수집된 데이터에 대하여 텍스트 클러스터링 분석 및 오피니언 마이닝을 통한 감성도 분석을 통하여 도시민들의 감성도를 수치화하였다. 그리고 두 키워드에 대하여 실험을 통하여 대도시일수록 시민들의 감정이 크게 변화하였음을 보였다.

잠재요인 모델에 기반하여 관객들의 선호도를 측정한 후에 영화를 추천하는 시스템이 제안되었고[8], 로짓 모형 형태의 선호도 모형을 이용한 스마트 주차 서비스가 제시되었다[9]. 그리고 생체신호인 PPG, GSR 신호 변화에 따라 구동되는 LED 감성 조명 시스템을 제안되었다[10].

III. 키워드 기반 선호도 분석 시스템 설계

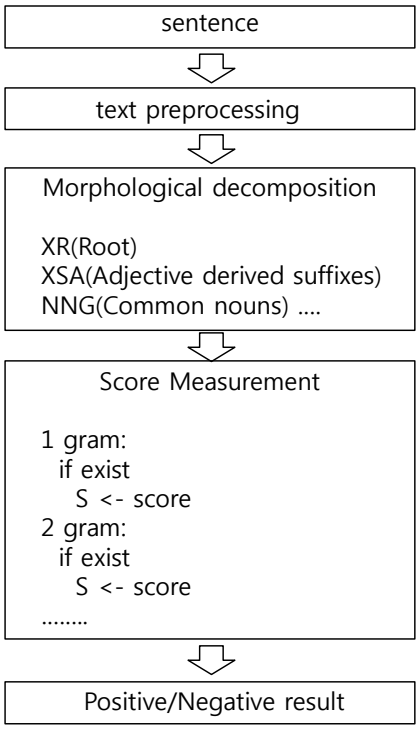

그림 1. N-gram 감정분석 절차
Fig. 1 N-gram sentiment analysis process

그림 1은 이 논문에서 사용될 N-gram 감정분석의 절차를 보여준다. 검사할 문장에 대하여 전처리 관정을 거친 후에 형태소별로 문장을 분해한다. 그리고 분해된 요소들을 사전에 구축된 감정 사전을 기반으로 N-gram에 따라 점수를 계산하여 해당 문장에 대한 긍정/부정 점수를 집계한다.

표 1. 문장에 대한 형태소 분해 예
Table. 1 Example of morphological decomposition for a sentence

morpheme	class
담백	어근(XR)
하	형용사 파생 접미사(XSA)
연출	보통명사(NNG)
이	주격 조사(JKS)
좋	형용사(VA)
았	사제 선어말 어미(EPT)
어요	평서형 종결 어미(EFN)

표 1은 문장에 대하여 형태소 분해를 한 예를 보여준다. 하나의 문장은 예와 같이 어근, 형용사 파생 접미사 등의 요소들로 분해된다. 표 2는 N-gram 감정분석 절차에 따라 평가된 문장의 부정/긍정 점수의 예를 보여준다. 표 2의 예와 같이 긍정적인 문장에 대해서는 긍정 점수가 높게 도출되고 부정적인 문장에 대해서는 부정 점수가 높게 도출된다.

표 2. 문장에 대한 감정 분석 점수 예
Table. 2 Example of sentiment analysis score for a sentence

sentence	POS	NEG
"담백한 연출이 좋았어요"	0.7854	0.2146
"싸늘하다. 가슴에 비수가 날아와 꽂힌다."	0.1397	0.8603
"큰일 났네. 모두 커피, 커피, 커피를 원해. 차는 잔뜩 남고 커피는 모자랄 지경이야!"	0.3577	0.4293

그림 2는 이 논문에서 제안하는 키워드 기반 선호도 분석 시스템의 처리 절차를 보여준다. 그림 2에서 보듯이 사용자가 긍정/부정을 분석하기 위한 키워드를 입력하면 입력 키워드와 관련된 키워드들을 추출한다. 그리고 추출된 키워드를 기반으로 웹의 검색 엔

진을 이용하여 텍스트 형태의 최근 웹 문서데이터들을 수집한다. 수집된 웹 문서데이터들은 데이터 전처리 과정을 거친 후에 N-gram 감정 분석 기법을 이용하여 문서 자료들에 대하여 긍정/부정을 계산한 후에 취합된 최종 결과를 사용자에게 제시한다.

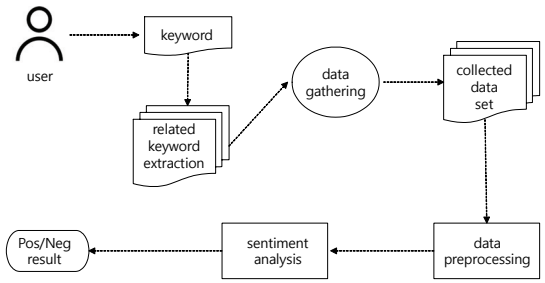

그림 2. 제안 시스템 처리 절차
Fig. 2 Proposed system processing flow

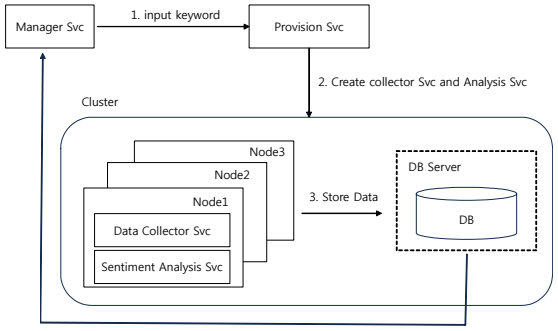

그림 3. 시스템 구성도
Fig. 3 System diagram

그림 3은 키워드 기반 선호도 분석 시스템 구성도를 보여준다. 그림 2의 시스템 설계에서 처리 부하가 가장 높은 부분이 데이터를 수집하는 모듈과 수집된 웹 문서데이터들에 대한 감정 분석을 수행하는 모듈이다. 이를 해결하기 위하여 이 논문에서는 마이크로서비스구조(Microservice Architecture)를 기반으로 시스템을 구성한다. 자료 수집 모듈과 감정 분석 모듈을 하나의 노드로 구성하고 다수의 노드를 동시 수행하여 자료 수집 및 감정 분석 처리 시간을 감소시킨다.

노드에서 수행되는 각 모듈은 컨테이너화한 서비스인 그림 4의 워커서비스(worker service) 형태로 구현되며 수행할 작업의 양에 따라 각 노드에서 생성(provisioning)되어 서비스를 수행한다. 각 워커서비스는 내부적으로 컨테이너서비스들로 구성되며 구성 요소는 내부 데이터베이스와 데이터 수집, 감정 극성 계산 컨테이너이다. 내부 서비스들은 gRPC를 통해 서로 통신하며 각 모듈 간의 결합도를 낮춘다.

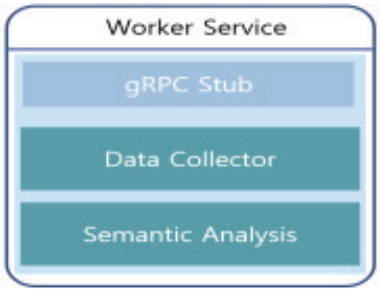

그림 4. 워커서비스 구조
Fig. 4 Worker service structure

표 3. Manager 클래스 명세
Table 3. Manager class specification

Manager	
Name	Description
CreateNewJob()	create job using a keyword
GetJob()	return status fo the job
DeleteJob()	delete job
GetJobList()	return the list of Job and status
GetArticles()	return articles collected in a job
UpdateJob()	modify the statis of the job

표 4. Provisond 클래스 명세
Table 4. Provisond class specification

Provisond	
Name	Description
CreateAnalyzer()	create analyzer
RemoveAnalyzer()	delete analyzer
CreateScraper()	create scraper
RemoveScraper()	delete scraper
UpdateWorker()	moduify the status of a worker

표 5. Scraped 클래스 명세
Table 5. Scraped class specification

Scraped	
Name	Description
GetArticles()	collect articles related with a keyword
MakeUniq()	remode replicated data in the collected articles
SendDoneMsg()	send the trigger of a job completion

키워드 기반 선호도 분석을 수행하기 위하여 Manager, Scraped, Provisiond, Analyzer로 크게 4가지의 서비스로 정의하였다. 각 서비스는 IDL(: Interface Defined Language)의 하나인 protobuf으로 정의하고 gRPC를 통해 각 서비스간에 데이터를 전달한다. 표 3, 표 4, 표 5는 각 서비스를 정의한 클래스 명세를 보여준다.

IV. 시스템 구현 및 결과

키워드 기반 선호도 분석 시스템의 구현을 위하여 ARM Cortex-A72(2GHz)와 8GB 램을 장착한 Broadcom BCM2711를 이용하여 SoC 클러스터로 구현하였다. 그리고 서비스 단위로 구현된 컨테이너의 관리를 위해 Docker Swarm을 사용하였다. 클러스터는 1개의 리더노드와 3개의 워커노드로 구성하였다.

제안 시스템은 처리 시간이 가장 많이 소요되는 웹 문서데이터 수집 시간과 감정 분석 수행 시간의 개선을 위하여 두 모듈을 하나의 노드에서 구성하고 여러 노드를 동시 수행하였다. 처리 시간 개선을 확인하기 위하여 47,146byte의 크기로 약 27,562자의 문자로 구성된 500개의 문장에 대하여 노드 1개를 사용한 수행 시간과 3개의 노드로 이루어진 클러스터를 사용한 수행 시간을 측정하였다. 표 6은 수행 시간의 비교 결과를 보여준다. 표 6에서 보듯이 싱글 노드에서 수행하였을 때 약 324초가 소요되었으며, 3개의 노드로 구성된 클러스터에서 수행하였을 때 약 109초가 소요되어 약 30%의 개선 효과를 보여주었다. 이는 다수의 노드로 구성한 클러스터를 사용하였을 때 데이터 통신 시간을 감안하더라도 작업을 수행하기 위한 노드의 개수만큼 처리 시간이 개선되는 것을 알 수 있다.

표 6. 자료수집 및 감정분석 소요 시간
Table 6. Processing time for data gathering and sentiment analysis

Node 1ea	Node 3ea
324sec	109sec

표 7. 카카오 서비스 장애에 대한 비교
Table 7. Comparison for kakao service failure

Kakao Service Error	Negative	Positive	Neutral
Gallup	59%	38%	3%
analysis result	52%	46%	2%

표 8. 대통령 키워드에 대한 비교
Table 8. Comparison for president keyword.

President	Negative	Positive	Neutral
Gallup	65%	27%	8%
analysis result	64%	35%	1%

표 7과 표 8은 2022년 10월 3주차에 2개의 키워드인 '카카오'와 '대통령'에 대하여 웹 문서데이터를 수집 후 선호도 분석을 실시한 결과와 여론 조사 기관인 갤럽(Gallup)사에서 해당 키워드에 대하여 여론 조사를 통해 도출한 결과의 비교를 보여준다. 2가지 키워드에 대하여 수집된 웹 문서데이터로부터 약 5일간 전처리 과정을 거친 5,800여 개의 문장과 6,055여 개의 문장을 분석하였다. 제안 시스템의 선호도 분석 결과와 갤럽사의 여론 조사 결과를 비교하였을 때 최소 1%에서 최대 8%의 차이를 나타내어 10% 이내의 오차율을 보여주고 있다. 8%의 오차율이 나타나는 이유는 검색 플랫폼의 특성과 감정 사전 도메인에 의한 것으로 판단된다. 특히 검색하고자 하는 대상에 따라 분석하여야 하는 도메인이 크게 달라진다. 예를 들어 정치에 관련된 키워드를 주었을 때 인터넷 기사가 많이 검색되지만, 사전의 도메인에 따라 대개 다른 극성 결과를 보여준다.

V. 결론 및 향후 연구

이 논문에서는 감정 측정을 이용한 선호도 분석 시스템을 제안하였다. 사용자가 입력한 키워드를 기반으로 관련 키워드를 생성한 후에 웹 문서데이터를 수집하였다. 그리고 수집된 웹 데이터들에 대하여 N-gram 기법을 이용하여 극성을 측정하였다. 또한 대용량 웹 데이터를 분석할 때 필요한 분석 시간을 줄이기 위하여 워커 서비스를 사용한 컨테이너 기반의 시스템을 설계 및 구현하였다. 향후 연구로는 GPT 기반의 트랜스포머 계열 감정 측정 기법을 사용한 선호도 분석 시스템을 설계하는 것이다.

감사의 글

본 연구는 2024년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구결과로 수행되었음 (2019-0-01817)

References

[1] Gallup. "Korea Daily Opmion", Gallup, 2022. 10.
[2] J. Lee, “Comparison of Sentiment Calssification Performance of for RNN and Transformer-Based Models on Korean Reviews,“ *Journal of the Korea Institute of Electronic Communication Sciences*, vol. 18, no. 4, 2023. 08, pp. 693-700.
[3] D. Kim, S. H. Lee, and J. H. Bong, “Artificial Intelligence for Assitance of Facial Expression Practice Using Emotion Classification,“ *Journal of the Korea Institute of Electronic Communication Sciences*, vol. 17, no. 6, 2022. 12, pp. 1137-1143.
[4] D. Chung, "A Case Study of Visualizing Emotions with Social Media Emotion Analysis : Focuesd on Media Art Cases," *Archives of Design Research*, vol. 35, no. 1, 2022. 02, pp. 237-257.
[5] H. Kim and J. Lee, "Sentiment Analysis Using Mixed Feature Vector combined with the Sentiment DictionaryH. Kim and J. Lee, "Sentiment Analysis Using Mixed Feature Vector combined with the Sentiment Dictionary Information," *J. of the Korean Institute of Intelligent System*, vol. 30, 2020. 12, pp. 494-499 Information", Journal of the Korean Insititue of Inteligent System, Vol.30, pp. 494-499, Dec. 2020.
[6] J. Jeong, J. Jang, and M. Moon, “Development of AI Speaker with Active Interaction Customized for the Elderly,“ *Journal of the Korea Institute of Electronic Communication Sciences*, vol. 15, no. 6, 2020. 12, pp. 1223-1229.
[7] M. Song and H. Yoo, "Citizen Sentiment Analysis of the Social Disaster by Using Opinion Mining," *J. of Korean Society for Geospatial Information Science*, vol. 25, no. 1, 2017. 04, pp. 37-46.
[8] C. Ma and K. Kim, "Movie Recommendation System based on Latent Factor Model," *Journal of the Korea Institute of Electronic Communication Sciences*, vol. 16, no. 1, 2021. 02, pp. 125-133.
[9] J. Jang, H. Lee, W. Lee, H. Kim and T. Kim, "A Study on Design Requirement for Smart Parking Services Considering User’s Stated Preferences," *Journal of the Korea Institute of Electronic Communication Sciences*, vol. 16, no. 6, 2021. 12, pp. 1279-1286.
[10] Y. Han and D. Kim, "Sensitivity Illumination System Using Biological Signal," *Journal of the Korea Institute of Electronic Communication Sciences*, vol. 9, no. 4, 2014. 04, pp. 499-507.

저자 소개

문희준(Hee-Jun Moon)

2016년 동서대학교 컴퓨터공학부 입학
※ 관심분야 : 인공지능, 딥러닝

김동현(Dong Hyun Kim)

1995년 부산대학교 컴퓨터공학과 졸업 (공학사)
1997년 부산대학교 대학원 컴퓨터공학과 졸업(공학석사)
2003년 부산대학교 대학원 컴퓨터공학과 졸업(공학박사)
2004년 ~ 현재 동서대학교 소프트웨어학과 교수
※ 관심분야 : 데이터베이스, 공간 데이터베이스, GIS, 센서데이터베이스